# CS540 Spring 2010: homework 5

This homework requires that you read the new "Regularized discriminant analysis" and "MAP estimation for EM" handouts, and download the latest version of PMTK (5feb2010).

## 1 Mode of an Inverse Wishart distribution

Show that

$$\arg\max \text{IW}(\boldsymbol{\Sigma}|\mathbf{S}_0, \nu_0) = \frac{\mathbf{S}_0}{\nu_0 + D + 1} \tag{1}$$

Hint: The proof of this is very similar to the derivation of the MLE of $\boldsymbol{\Sigma}$ on p188 of the book.

## 2 Derivation of the mode of an NIW distribuiton

Consider the distribution $\text{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{m}_0, \kappa_0, \mathbf{S}_0, \nu_0)$. Show that the joint mode of this is given by

$$\hat{\boldsymbol{\mu}} = \frac{\mathbf{m}_0}{\kappa_0} \tag{2}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{\mathbf{S}_0}{\nu_0 + D + 2} \tag{3}$$

Hint: see section 30.11 of the book for some helpful identities from matrix calculus.

## 3 Derivation of the NIW posterior

Show that the posterior for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ under a Gaussian likelihood using a conjugate NIW prior has the following form:

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D}) = \text{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{m}_n, \kappa_n, \nu_n, \mathbf{S}_n) \tag{4}$$

$$\mathbf{m}_n = \frac{\kappa_0 \mathbf{m}_0 + N\overline{\mathbf{x}}}{\kappa_n} = \frac{\kappa_0}{\kappa_0 + N}\mathbf{m}_0 + \frac{N}{\kappa_0 + N}\overline{\mathbf{x}} \tag{5}$$

$$\kappa_n = \kappa_0 + N \tag{6}$$

$$\nu_n = \nu_0 + N \tag{7}$$

$$\mathbf{S}_n = \mathbf{S}_0 + \mathbf{S}_{\overline{x}} + \frac{\kappa_0 N}{\kappa_0 + N}(\overline{\mathbf{x}} - \mathbf{m}_0)(\overline{\mathbf{x}} - \mathbf{m}_0)^T \tag{8}$$

Hint: see the appendix of [FR05] for some very helpful algebraic identities. (Essentially the answer is already there, but it is somewhat buried by a mass of notation.) I have put this paper on the class web page for convenience.

## 4 The Wishart distribution and friends

An alternative to using an inverse Wishart distribution on the covariance matrix is to use a Wishart distribution on the precision matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$. This is defined as

$$\text{Wi}(\boldsymbol{\Lambda}|\mathbf{S}_0, \nu_0) \propto |\boldsymbol{\Lambda}|^{(\nu_0 - D - 1)/2} \exp\left(-\frac{1}{2}\text{tr}(\boldsymbol{\Lambda}\mathbf{S}_0^{-1})\right) \tag{9}$$

Similarly, one can define

$$\text{NW}(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{m}_0, \eta_0, \nu_0, \mathbf{S}_0) \stackrel{\text{def}}{=} \mathcal{N}(\boldsymbol{\mu}|\mathbf{m}_0, (\eta_0\boldsymbol{\Lambda})^{-1}) \times \text{Wi}(\boldsymbol{\Lambda}|\mathbf{S}_0, \nu_0) \tag{10}$$

This formulation is widely used in machine learning, whereas the NIW formulation is more popular in statistics. Hence it is useful to have results in both forms.

1. Derive the mode of a Wishart distribution.
2. Derive the joint mode of an NW distribution.
3. Derive the posterior for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ assuming a NW prior and a Gaussian likelihood.

# 5 MAP EM for GMMs

Implement MAP estimation for GMMs using a NIW or NW prior. Set the hyper-parameters as described in the handout. Plot the log likelihood plus log prior vs iteration, and check it increases monotonically (on any dataset you choose, e.g., old faithful).
Then create a synthetic data set on which ML estimation fails (due to a singular $\hat{\boldsymbol{\Sigma}}_k$) but MAP estimation succeeds. (Remember to set your random number seed to ensure reproducibility.)
Bonus points: make a plot of the fraction of times ML and MAP fails (i.e., have numerical problems) vs $D$, for various $K$ and fixed $N$, averaging over multiple random restarts, as well as multiple random data sets. (I think you'll find that MAP never fails, whereas MLE becomes increasingly prone to fail as either $D$ or $K$ increase.)
Turn in your code and plot.

# 6 EM for mixtures of Students

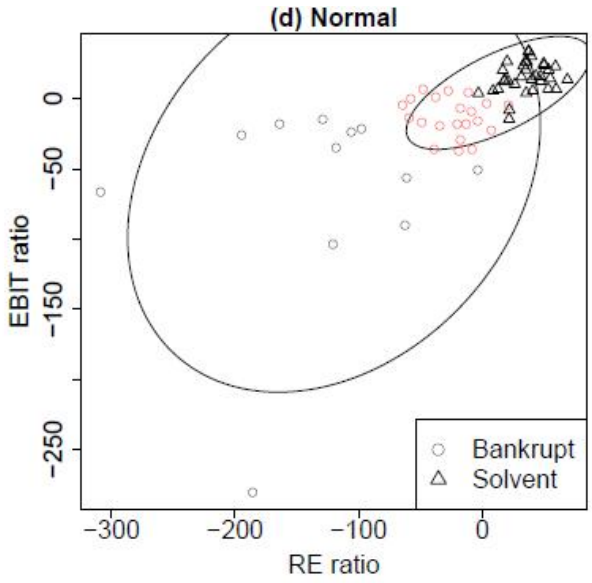
Implement the EM algorithm for finding the MLE of a mixture of Student distributions. Estimate $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and $\nu_k$ for each cluster $k$. Call the function `mixStudentFitEm`.
Apply `mixStudentFitEm` and `mixGaussFitEm` (part of pmtk) to the $N = 66$,$D = 2$ bankruptcy data set on the class website, using $K = 2$ clusters. The first column of this file specifies if a firm went bankrupt or not; the second column specifies the ratio of retained earnings (RE) to total assets; and the third column specifies the raio of earnings before itnerests and taxes (EBIT) to total assets. Ignore the first column for fitting purposes (i.e., do unsupervised clustering).
Now try to reproduce Figure 1. In particular, for each model, plot the 90% level sets of each component. Superimpose the data on the plot, using the true labels to specify the type of symbol (circle for bankrupt, triangle for solvent). Then perform a hard clustering of each data point, and if the estimated cluster assignment $z_i$ does not equal the true class label $y_i$, count it as an error and color the symbol red. (Try both possible interpretations of the latent labels, and pick the one with the lowest overall error rate.) What error rates do you get for each model on the training set?[1] Turn in your code, plots and error rates.

# References

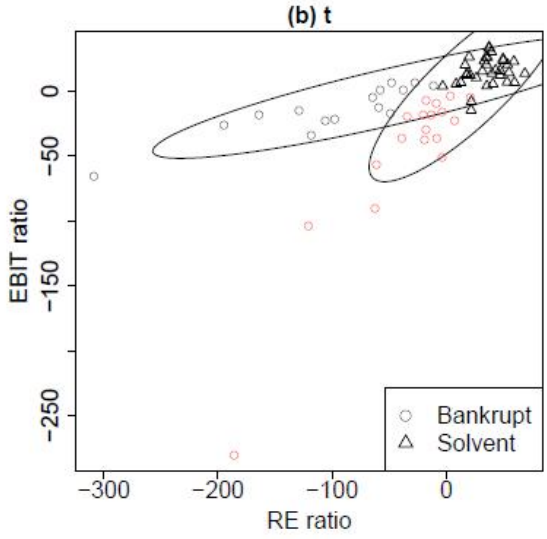

[FR05] C. Fraley and A. Raftery. Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. Technical report, U. Washington, 2005.

[Lo09] C. H. Lo. *Statistical methods for high throughput genomics*. PhD thesis, UBC, 2009.


[1] Kenneth Lo found that the mixture of Gaussians made 21 errors, and that the mixture of Students made 18 errors. You should be able to reproduce these numbers. (The reason the Gaussian model is worse is because there are a few outliers in the bankrupt class, which adversely effects the estimate of the covariance matrix for that cluster.)




(a)




(b)

*Figure 1:* (a) Mixture of 2 Gaussians fit to bankrupty data. (b) Mixture of 2 Students fit to bankrupty data. Source: Figure 3.3 of [Lo09].